CLAIMS

What is claimed is:

1. A method, in a network comprising a primary server and at least one offload server, for dynamic offloading of processing requests from said primary server to said at least one offload server, the method comprising the steps of:

determining a load on said primary server;

if the load on said primary server is less than a first threshold, serving processing requests at said primary server; and

if the load on said primary server exceeds said first threshold, then offloading at least a portion of said processing requests to said at least one offload server.

2. The method of claim 1 wherein said load comprises bandwidth utilization and said first threshold is a network bandwidth utilization of said primary server.

3. The method of claim 1 wherein the said load comprises CPU utilization and said first threshold is a CPU utilization of said primary server.

4. The method of claim 1 wherein serving the processing requests at said primary server includes returning a page to a user wherein all the embedded objects in the page have links to said primary server; and

offloading at least a portion of the processing requests to at least one offload server includes serving a base page at said primary server in which the links for embedded objects point to said offload server.

5. The method of claim 1 wherein offloading at least a portion of the processing requests to said at least one offload server includes routing an incoming Web request to a selected offload server.

6. The method of claim 1 and further including the step of, if the processing load on said primary server exceeds a second threshold, throttling at least one processing request.

7. The method of claim 6 wherein throttling at least one processing request includes returning a page to a user indicating that a server is overloaded

8. The method of claim 6 wherein throttling at least one processing requests includes dropping the at least one processing request without returning any information to a user.

9. The method of claim 6 wherein throttling at least one processing request includes returning a page to a user indicating that a server is overloaded if said load exceeds said second threshold, and dropping said at least one processing request if said load exceeds a third threshold.

10. The method of claim 1 wherein the determination of which of said at least one offload server that at least one processing request is to be offloaded to is based on one or more of a group including a client identity, a client gateway (IP) address, a price of the offload service, or a current or previous load on the at least one offload server.

11. A network apparatus comprising a primary server and at least one offload server connected by an IP-based network, for dynamic offloading of processing requests from said primary server to said at least one offload server, the network apparatus comprising:

a load controller connected between said network and said primary server;

a memory connected to said load controller and including data and control instructions for operating said primary server to perform the steps of:

determining a load on said primary server;

if said load on said primary server is less than a first threshold, serving processing requests at said primary server; and

if said load on said primary server exceeds said first threshold, then offloading at least a portion of said processing requests to said at least one offload server.

12. The apparatus of claim 11 wherein said load comprises network bandwidth and said first threshold is a network bandwidth utilization of said primary server.

13. The apparatus of claim 11 wherein said load comprises CPU utilization and said first threshold is a CPU utilization of said primary server.

14. The apparatus of claim 11 wherein serving the processing requests at said primary server includes returning a page to a user wherein all the embedded objects in the page have links to said primary server; and

offloading at least a portion of the processing requests to at least one offload server includes serving a base page at said primary server in which the links for embedded objects point to said offload server.

15. The apparatus of claim 11 wherein offloading at least a portion of the processing requests to said at least one offload server includes routing an incoming Web request to a selected offload server.

16. The apparatus of claim 11 and further including, if the processing load on said primary server exceeds a second threshold, throttling at least one processing request.

17. The apparatus of claim 16 wherein throttling at least one processing request includes returning a page to a user indicating that a server is overloaded

18. The apparatus of claim 16 wherein throttling at least one processing requests includes dropping the at least one processing request without returning any information to a user.

19. The apparatus of claim 16 wherein throttling at least one processing request includes returning a page to a user indicating that a server is overloaded if the primary server load exceeds said second threshold, and dropping the at least one processing request if said primary server load exceeds a third threshold.

20. The apparatus of claim 11 wherein the determination of which of said at least one offload server that at least one processing request is to be offloaded to is based on one or more of a group including a client identity, a client gateway (IP) address, a price of the offload service, or a current or previous load on the at least one offload server.

21. A system, including an IP network comprising a primary server and at least one offload server, for dynamic offloading of processing requests from said primary server to said at least one offload server, the system comprising:

means for determining a load on said primary server;

means for, if said load on said primary server is less than a first threshold, serving the processing requests at said primary server; and

means for, if said load on said primary server exceeds said first threshold, offloading at least a portion of said processing requests to said at least one offload server.

22. A program product including control instructions for controlling the operation of a computer, said program product operative with said control instructions to operate said computer in an IP-based network comprising a primary server and at least one offload server to dynamically offload processing requests from said primary server to said at least one offload server, the computer operative with said control instructions to perform the steps of:

determining a load on said primary server;

if said load on said primary server is less than a first threshold, serving the processing requests at said primary server; and

if said load on said primary server exceeds said first threshold, then offloading at least a portion of said processing requests to said at least one offload server.

23. A system for allocating processing requirements on a network between a primary server and an offload server, comprising:

a load controller connected to said network for receiving processing requests from clients on said network and allocating said processing requests between said primary and offload servers;

a memory connected to said load controller and storing threshold data and control software for analyzing said threshold data and operating said load controller;

said load controller operative with the threshold data and control software to perform the steps of

periodically evaluating said processing requests to determine a load on said primary server;

if said load exceeds a first threshold, for a predetermined period of time directing at least one processing request to said offload server; and

if said load does not exceed said first threshold, directing said processing requests to said primary server.

24. The system of claim 23 wherein said load is network bandwidth and said first threshold is a measure of the network bandwidth utilization of the primary server.

25. The system of claim 23 wherein said load is CPU utilization and said first threshold is a measure of the CPU utilization of the primary server.

26. The system of claim 23 wherein directing said processing requests to said primary server further includes returning a page to a user wherein all the embedded objects in the page have links to said primary server; and

directing at least one processing request to said offload server further includes serving a base page at said primary server in which the links for embedded objects point to said offload server.

27. The system of claim 23 wherein directing at least one processing request to said offload server further includes routing an incoming Web request to a selected offload server.

28. The system of claim 23 and further including, if said load exceeds a second threshold, throttling at least one processing request by returning a page to a user indicating that a server is overloaded

29. The system of claim 23 and further including, if said processing load exceeds a second threshold, dropping said at least one processing request without returning any information to a user.

30. The system of claim 23 and further including throttling at least one processing request by returning a page to a user indicating that said primary server is overloaded if said primary server load exceeds a second threshold, and dropping the at least one processing request if said primary server load exceeds a third threshold.

31. The system of claim 23 wherein said offload server includes a plurality of offload servers and further including determining which of said plurality of offload servers that at least a portion of one processing request is to be offloaded to is based on one or more of a group including a client identity, a client gateway (IP) address, a price of the offload service, or a current or previous load on the at least one offload server.

32. A method for allocating processing requirements on an IP network between a primary server and an offload server, comprising:

periodically evaluating processing requests to determine a load on said primary server;

if said load exceeds a first threshold, for a predetermined period of time directing at least one processing request to said offload server; and

if said processing load does not exceed said first threshold, directing said processing requests to said primary server.

33. The method of claim 32 wherein said load comprises network bandwidth and said first threshold is a measure of the network bandwidth utilization of said primary server.

34. The method of claim 32 wherein said load comprises CPU utilization and said first threshold is a measure of the CPU utilization of said primary server.

35. The method of claim 32 wherein directing said processing requests to said primary server further includes returning a page to a user wherein all the embedded objects in the page have links to said primary server; and

directing at least one processing request to said offload server further includes serving a base page at said primary server in which the links for embedded objects point to said offload server.

36. The method of claim 32 wherein directing at least one processing request to said offload server further includes routing an incoming Web request to a selected offload server.

37. The method of claim 32 and further including, if said load exceeds a second threshold, throttling at least one processing request by returning a page to a user indicating that a server is overloaded

38. The method of claim 32 and further including, if said load exceeds a second threshold, dropping the at least one processing request without returning any information to a user.

39. The method of claim 32 and further including throttling at least one processing request by returning a page to a user indicating that the primary server is overloaded if the primary server load exceeds a second threshold, and dropping the at least one processing request if the primary server load exceeds a third threshold.

40. The method of claim 32 wherein said offload server includes a plurality of offload servers and further including determining which of said plurality of offload servers that at least a portion of one processing request is to be offloaded to is based on one or more of a group including a client identity, a client gateway (IP) address, a price of the offload service, or a current or previous load on the at least one offload server.

41. A system for allocating processing requirements on an IP network between a primary server and an offload server, comprising:

means for periodically evaluating processing requests to determine a load on said primary server;

means for, if said load exceeds a first threshold, for a predetermined period of time directing at least one processing request to said offload server; and

means for, if said processing load does not exceed said first threshold, directing said processing requests to said primary server.

42. A program product including control instructions for controlling the operation of a computer, said program product operative with said control instructions to operate said computer in an IP-based network comprising a primary server and at least one offload server to dynamically offload processing requests from said primary server to said at least one offload server, the computer operative with said control instructions to perform the steps of:

periodically evaluating processing requests to determine a load on said primary server;

if said load exceeds a first threshold, for a predetermined period of time directing at least one processing request to said offload server; and

if said load does not exceed said first threshold, directing said processing requests to said primary server.